II

Amino Acid Alignments

The amino acid sequence alignments that follow were generated by the PIMA program (“pattern induced multiple alignment”; R.F.Smith and T.F.Smith, *PNAS* **87**:118, 1990 and *Protein Engineering* **5**:35, 1992). The alignments were also subject to some hand-editing performed with MASE (D.V. Faulkner and J. Jurka, *TIBS* **13**:321, 1988). For detailed information concerning these UNIX tools, contact Dr. Randall Smith, Institute of Molecular Genetics, Baylor College of Medicine, Houston, Texas, 713-798-4735.

With few exceptions (most notably with HIV-1 Env), only full-length protein sequences have been included in this compilation. Tables giving the basic information about each sequence contained in these alignments—locus name, accession number, author, and journal— precede each amino acid alignment; these tables also appear in Part I. Beginning in 1995, common names rather than GenBank Locus names, are presented in the alignments. Both common names and Locus names may differ from names in the literature, but the accession numbers will be universal identifiers.

Eight sequence subtypes have been identified for HIV-1s (see Part I and Part III) and five sequence subtypes exist for HIV-2s. Sequence subtypes have been defined by “cladistic” criteria applied to nucleotide sequences; it remains to be seen whether the amino acid sequence subtypes inferred from that classification are valid, or whether a “phenetic” classification would be preferable.

Mosaic (hybrid) sequences that may have resulted from recombinational events (see Part III) are sometimes aligned under a category designated U. At other times, they may be aligned under the consensus sequence that they best match, although they did not contribute to the constitution of that consensus; in this latter case, sequences analyzed to be mosaic, for example the HIV-1 MAL sequence, may have prefixes such as “AD_” that indicate the hybrid character.

The reference sequences for the alignments are mixed case consensus sequences in which upper case letters refer to amino acid residues which are conserved 100% and lower case letters represent amino acid residues conserved in at least 50% of the sequences. The symbol “?” indicates no consensus at a position. Consensus sequences have been generated for each of the defined subtypes (see Parts I and III), and these are presented both with the grand alignments and in alignment to one another. The user should keep in mind that these subtypes have been “cladistically” defined, not “phenetically” defined (the number of phenotypes remains to be discovered).

Within the sequences, “-” is used to indicate residues conserved with respect to the reference sequence and “.” represents actual gaps. The symbol “$” indicates a stop codon and the symbol “#” indicates a frameshift or untranslatable situation. Blank spaces within the alignment indicate lack of sequence information over that region. Annotation of the *env* amino acid sequences utilizes “*” for conserved cysteine residues and “ˆˆˆ” for potential N-linked glycosylation sites.

At the risk of relaxing standards of annotation, we have elected to annotate features for which some evidence supports a particular role for a motif—nls for nuclear localization signal, for example. The authority for these additions can be found for the most part in section III curatorial comments (see the Gag entry from 1994 or the Vpr entry from 1995, for example). We have revised in this issue the annotation pertaining to the CD4-binding of the Env gp120, with the following rationale provided by John Moore, Aaron Diamond AIDS Research Center, New York:

> The original designation of the ‘Lasky site’ as “the CD4 binding site” on gp120 was based principally on two observations: 1) Deletion of a segment of the C4 domain of gp120 did not prevent formation of a folded molecule, but did prevent it binding to CD4; 2) a peptide-reactive Mab recognizing the same area of C4 blocked gp120 binding to CD4.

The observations are correct, and unchallenged to this day. The inference drawn from them was also logical at the time. However, it is now considered simplistic to call a simple C4 continuous region (i.e., `RIKQIINMWQEVGKAMYAPPISGQIR`) the CD4-binding site. There are three reasons for this:

a) Mutation of almost all the above residues has little or no effect on CD4 binding; the exception is Trp-427, whose mutation completely abolishes CD4 binding. This residue is a prime candidate for a CD4 contact residue. However, mutation of nearby residues 423, 426, 428, 429, 430, 432, 433, 435, 438 has no significant effect (Sodroski's work).

b) Mutations elsewhere in gp120 are just as disruptive to CD4 binding as ones at Trp-427. The other critical residues include 368-D, 370-E and 457-D. Changes at 262-N also disrupt CD4 binding, but probably because of gp120 misfolding (Sodroski's work).

c) Other MAbs to gp120 block CD4 binding without either binding to the above C4 peptide or, in some cases, being sensitive to amino acid substitutions in C4.

Overall, the present view of the CD4 binding site is that it is a discontinuous structure formed by the juxtaposition of residues scattered among the gp120 sequence, but probably concentrated in the C-terminal half. One residue in C4 (W-427) probably contributes to the formation of the site, but so do others.

PART II Amino Acid Sequences and their Variation